MP 5.7 A 600MHz IA-32 Microprocessor with Enhanced Data Streaming for Graphics and Video

Stephen Fischer, Ramesh Senthinathan, Hamid Rangchi, Hadi Yazdanmehr

Intel Corporation, Folsom, CA

This Intel architecture microprocessor (Pentium®III) implements 70 additional instructions to further accelerate the performance of data-streaming applications including 3D graphics and video encode/decode. This processor is enhanced by addition of these instructions along with circuit improvements for higher clock frequency [2]. The 10.17x12.10mm^2 die contains 9.5M transistors and is in a CMOS 5-layer metal 0.25µ process in an OLGA package with C4 interconnect technology. It has an operating range of 1.4V to 2.2V and is currently running up to 600MHz.

The Pentium®III processor implements 53 new instructions and an 8-entry register file to support a new packed floating point data type. Each architectural register is 128b wide with four independent single-precision elements packed in parallel. Floating-point arithmetic modes are supported for full compliance with IEEE 754 standard in rounding, and for streamlining multimedia applications where under-flowed results are forced to a zero result.

The multiply and divide-related packed floating point operations are supported by reuse of the existing hardware. The multiplier is a fully pipelined Wallace tree based on a radix-8 Booth encoding with the primary change being the addition of 6 more Booth encoders. For the divider, the SRT 2b/cycle algorithm is used with an additional QPLA for quotient prediction on the second packed data element. The existing rounder, capable of handling the wider precision IEEE 754 formats, is similarly retrofitted to operate on independent packed data values by expanding the mantissa incrementer to operate on the exponent field for packed data as well. To sustain a throughput of two packed floating point add-multiply vector operations per cycle, a separate packed floating point adder is implemented on a different execution port from the multiplier. This is a fully pipelined 32b Kogge-Stone dynamic adder using an 8b grouping of the carry look ahead circuitry and leading zero anticipation logic during re-normalization, with 3-cycle latency 1-cycle throughput including the data adjustment through a separate dedicated rounder (Figure 5.7.1). Masked-mode numeric exception responses, provided in hardware for packed floating-point operations, minimize microcode intervention.

The SIMD unit in the Pentium®III processor responsible for executing MMX™ technology instructions is enhanced to accommodate 12 new packed-integer instructions [1]. New min and max functions are supported through the adder by using the carry-out to bypass one of the inputs to write-back. SIMD adder changes are avoided by conditionally switching the source inputs based on the min versus max operation. A sum of absolute differences instruction is accommodated within the existing adder and multiplier by breaking up the function into 3 micro-operations. The source operands are first subtracted with a resulting carry flag and data value per packed element. The second micro-operation performs an absolute value through the use of the carry to conditionally invert the source. The third micro-operation performs a horizontal add function by taking advantage of the dual Wallace trees within the SIMD multiplier. As shown by Figure 5.7.2, source data is inserted into the Wallace trees to line up vertically (with the rest of the bits zeroed), so the resulting intermediate sums again are combined utilizing the separate adder that exists for the PMADDWD instruction. These changes to existing SIMD hardware increase the area of this unit by less than 2% and yet enable the Katmai processor to support real-time MPEG II encode at 30frames/s.

A problem in achieving high write data bandwidth in typical data streaming applications is the non-temporal or non-caching nature of the target data, which does not benefit as much from traditional caches as typical integer desktop applications. Through the addition of new non-temporal store instructions and expanded write-combining buffer capabilities, outgoing write data bandwidth are sustained at 1.066GB/s with a 133MHz external bus. The impact of high read latencies, attributable to non-temporal data, is minimized by enabling the overlap of other meaningful code through the support for 4 new data prefetch instructions. Pipelined stalls are reduced by decoupling the availability of a fill buffer and completion of the prefetch from the retirement of the instruction. Outstanding prefetch operations (and other loads) are tracked in a load buffer that marks a prefetch complete and ready for retirement once the operation is dispatched in the memory unit. For maximum die area efficiency, write gathering, cache line fills, and prefetch operations use the same array of buffers.

To achieve Pentium®III processor operating frequency, numerous speed paths needed to be optimized. One category relates to the use of domino tag comparators incorporated in various caches and buffers across the micro-architecture comprised of 8T XOR cells. An XOR gate shown in Figure 5.7.3 outperforms previous implementations on this processor in terms of delay, contention, back-writing, and reduced cross-capacitance impact.

Primary- and secondary-stage process, voltage, and temperature (PVT) compensated drivers meet 133MHz front-side bus timing requirements. The back-side bus protocol incorporates a source-synchronous clocking scheme for support with both external vendors as well as in-house custom synchronous SRAMs. To overcome clock jitter limitations on the back-side bus at full-speed, the receivers are designed with phase-borrowing de-skewed latches.

To achieve performance goals above 500MHz in the presence of switching noise, guidelines are established for the placement of decoupling capacitors with respect to density and proximity to high-speed switching circuits. A methodology checks conformance to density rules while locality rules are guaranteed by requiring that decoupling capacitors not be placed more than 100µm away from a designated switching element. This distance is empirically derived by running a detailed parasitic extractor through design switching current/noise profile curves (Figure 5.7.4). These improvements are expected to produce an even better return for mobile implementations, which use a lower supply voltage.

Global clocking is supported with a dual spine clock tree. A delay lock line (DLL) balances the two main clock spines. The DLL design maintains a lower global clock skew by dynamically compensating forPVT variations. On-die-clock (ODC) shrink/stretch circuitry is incorporated into the clock unit for speed-path related debug (Figure 5.7.5). The ODC circuit operates up to 700MHz with the ability to shrink or stretch a single clock edge by as much as 112ps in 17ps steps as controlled through the test access port.

Acknowledgments:
The authors thank H. Dang, J. Delgross, M. Gley, M. Jahan, K. Kong, H. Lai, E. Lilya, P. Modali, M. Nazareth, S. Palanca, S. Qawami, J. Reese, A. Sama, S. Siers, and T. Woldeyes for contributions.

References:
[1] Choudhury, M. et. al., "A 300MHz CMOS Microprocessor with Multi-Media Technology", ISSCC Digest of Technical Papers, pp 170-171, Feb., 1997.
[2] Schutz, J., et al., "A 450MHz IA32 P6 Family Microprocessor", ISSCC Digest of Technical Papers, pp. 236-237, Feb., 1998,

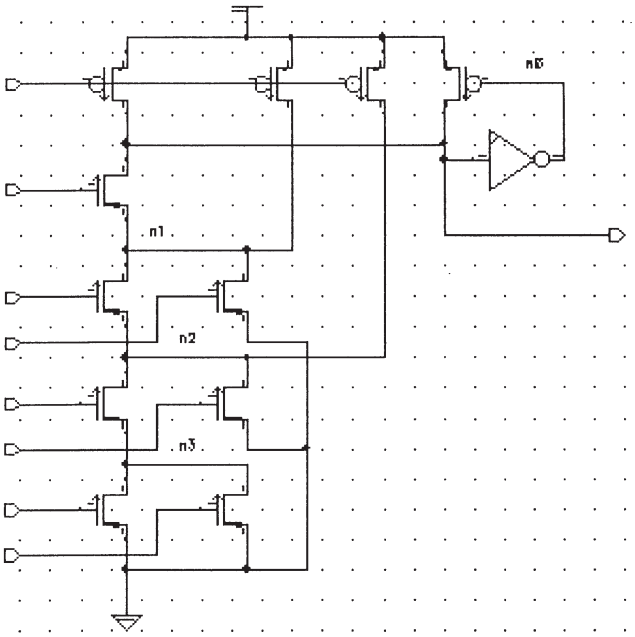

Figure 5.7.1: Carry look-ahead generator domino.

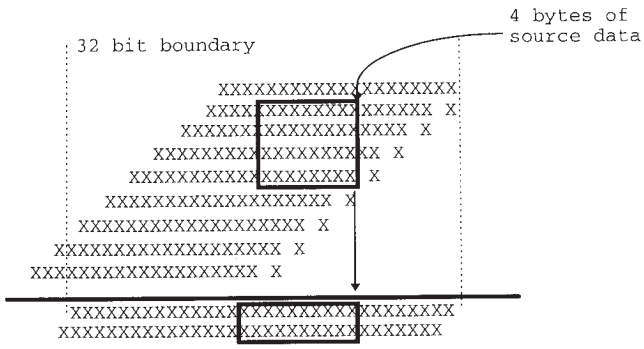

Figure 5.7.2: Data insertion in Wallace tree.

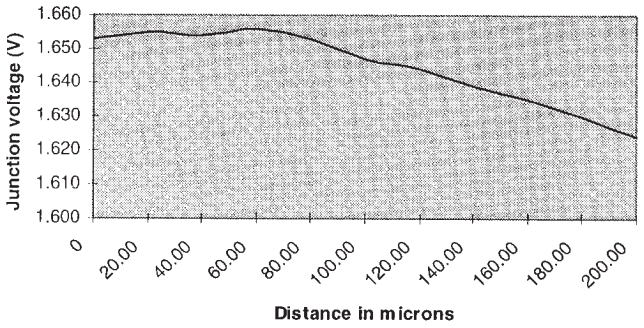

Figure 5.7.4: Junction voltage vs de-capdistance.

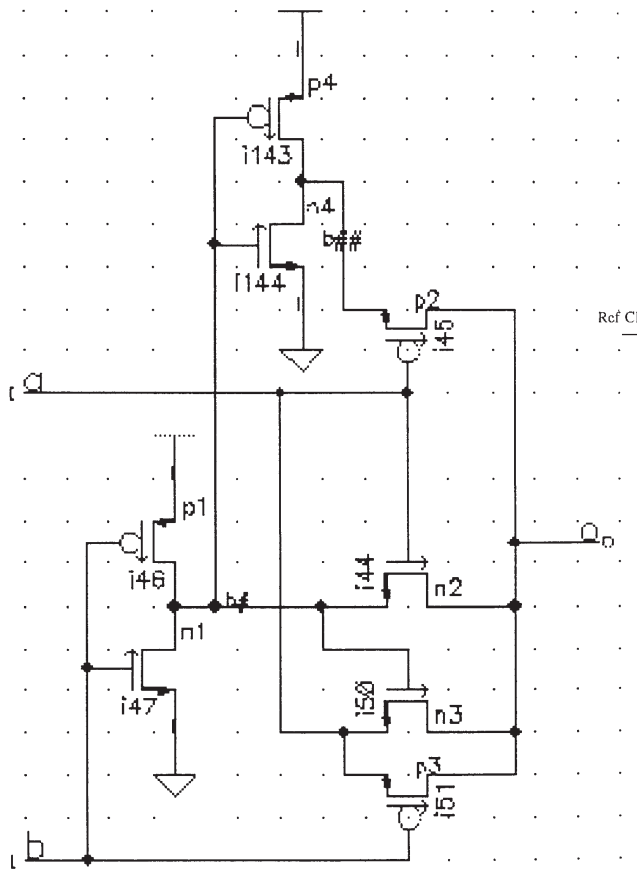

Figure 5.7.3: 8T XOR gate.

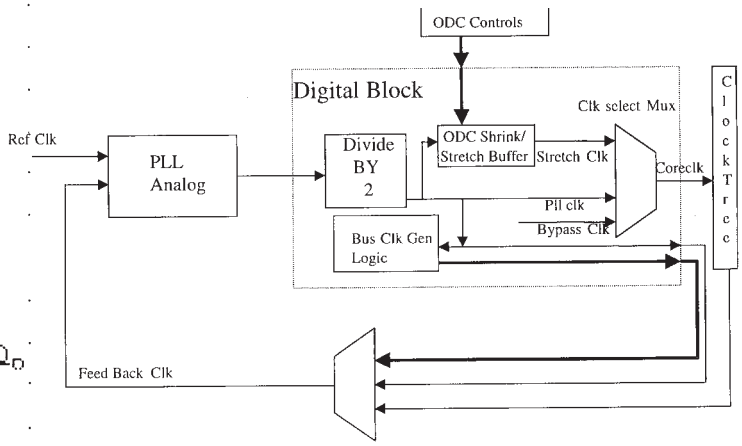

Figure 5.7.5: On-die-clock shrink/stretch functional block.

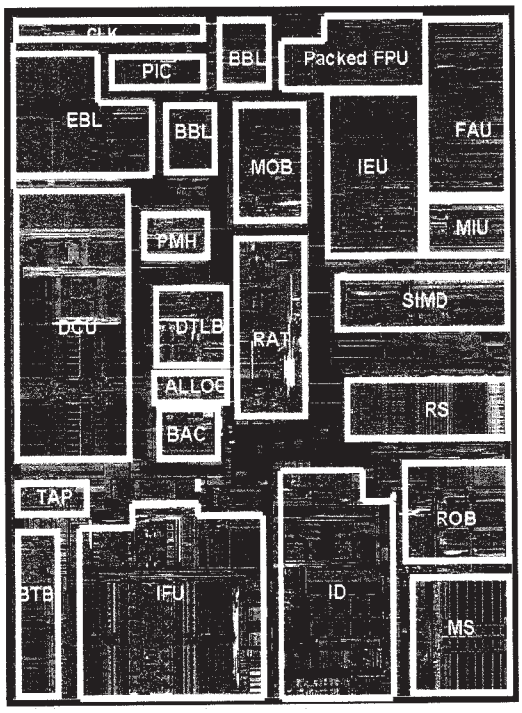

Figure 5.7.6: Die micrograph.

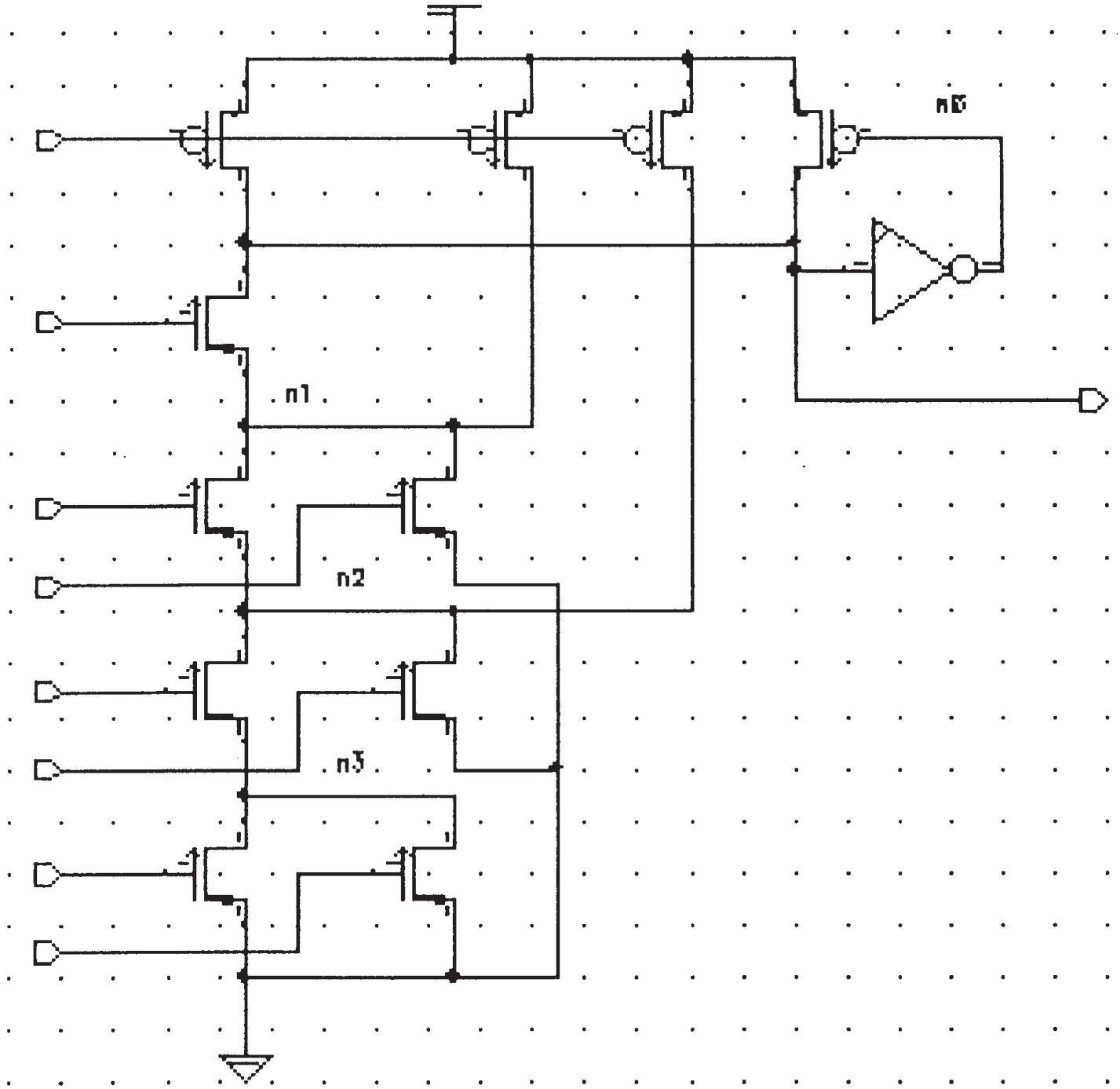

Figure 5.7.1: Carry look-ahead generator domino.

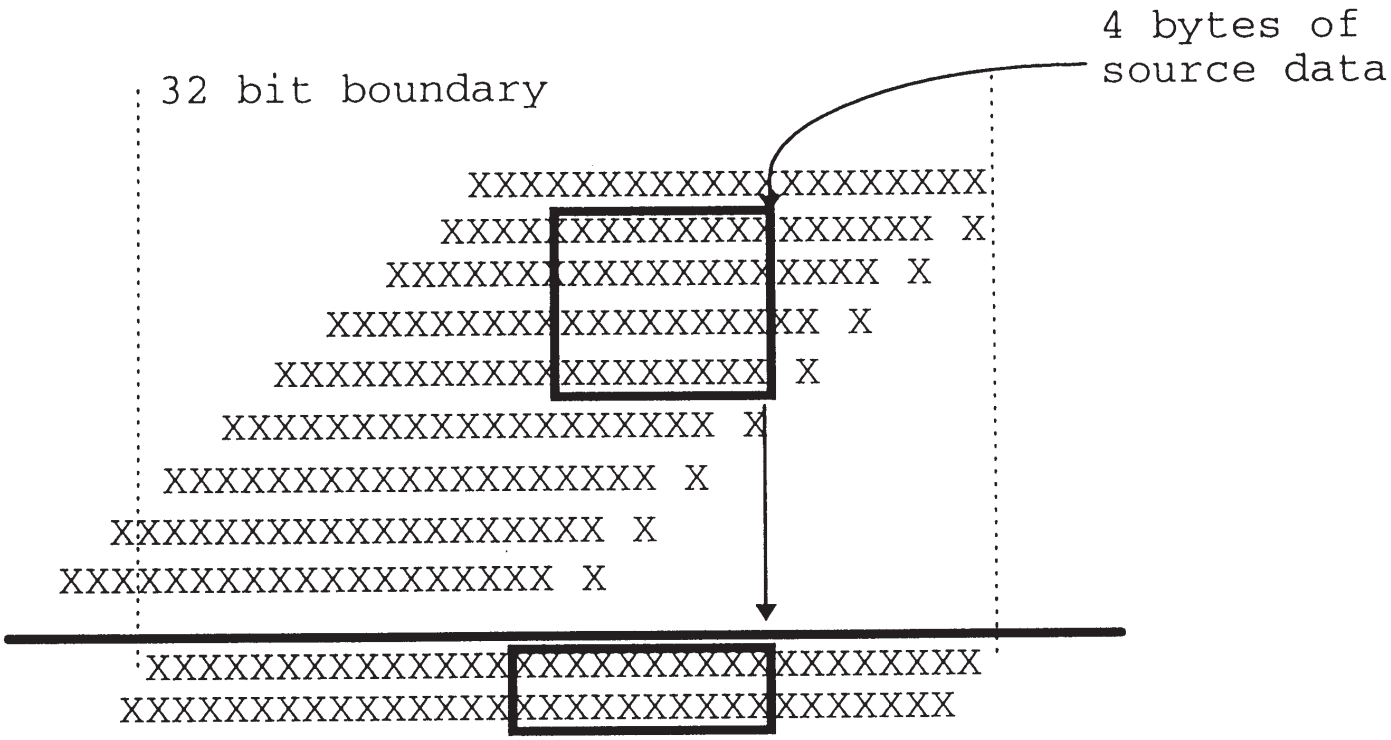

Figure 5.7.2: Data insertion in Wallace tree.

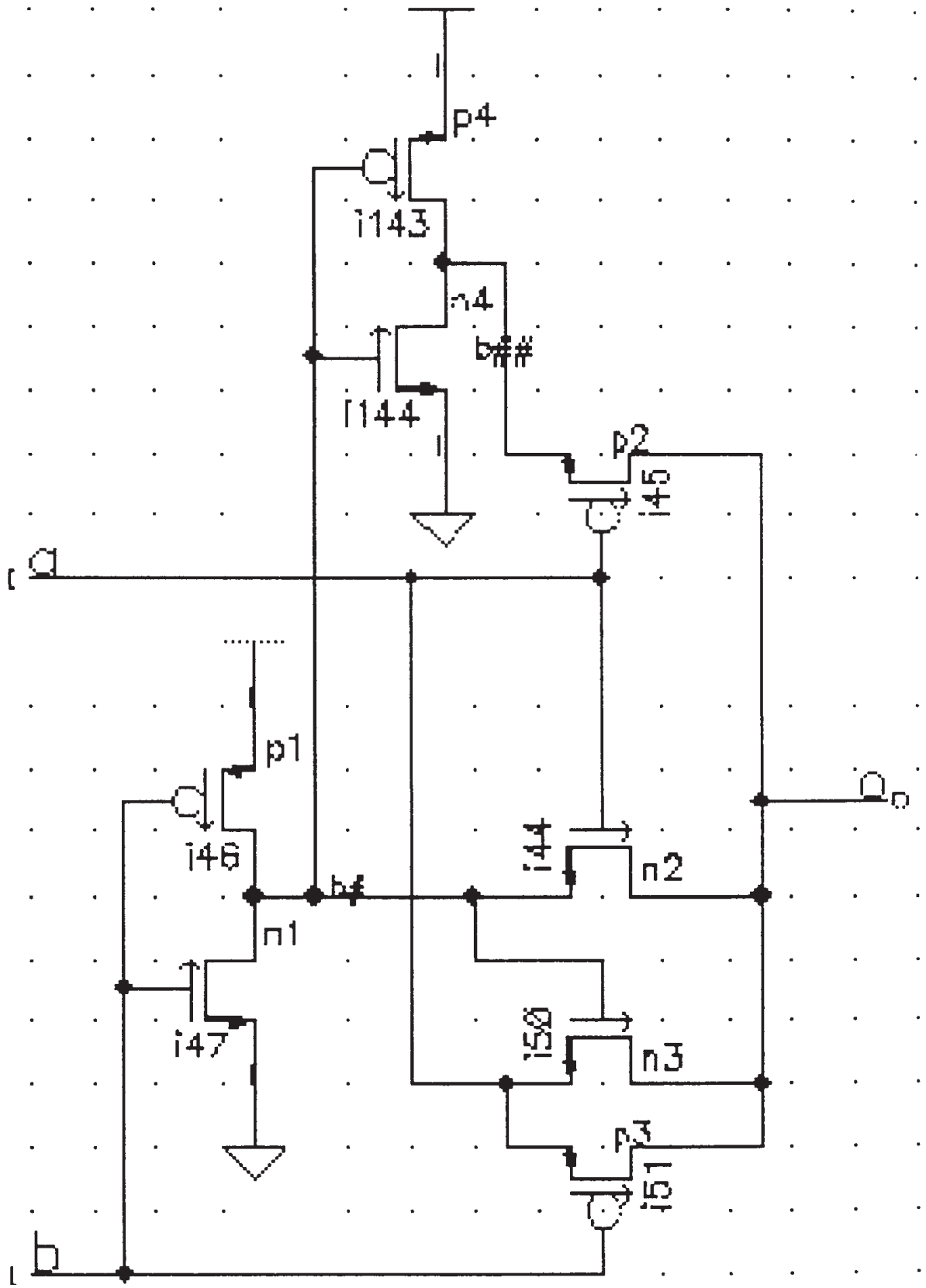

Figure 5.7.3: 8T XOR gate.

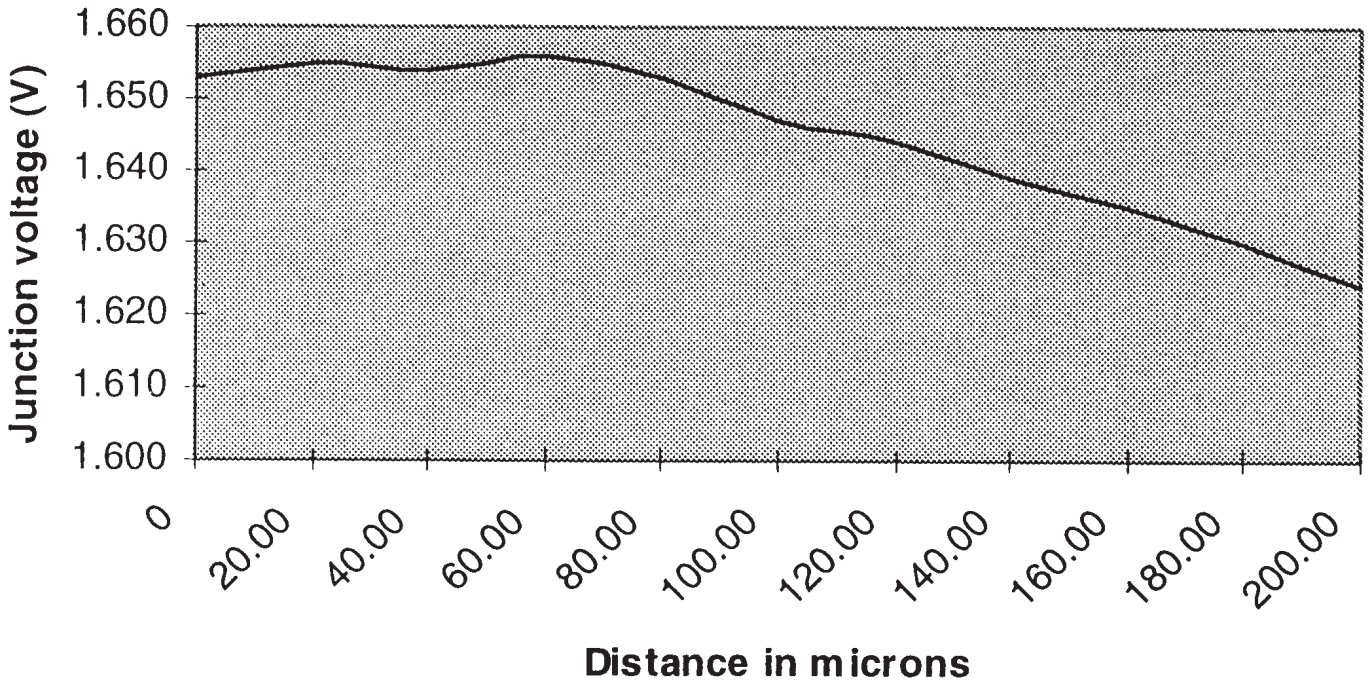

Figure 5.7.4: Junction voltage vs de-cap distance.

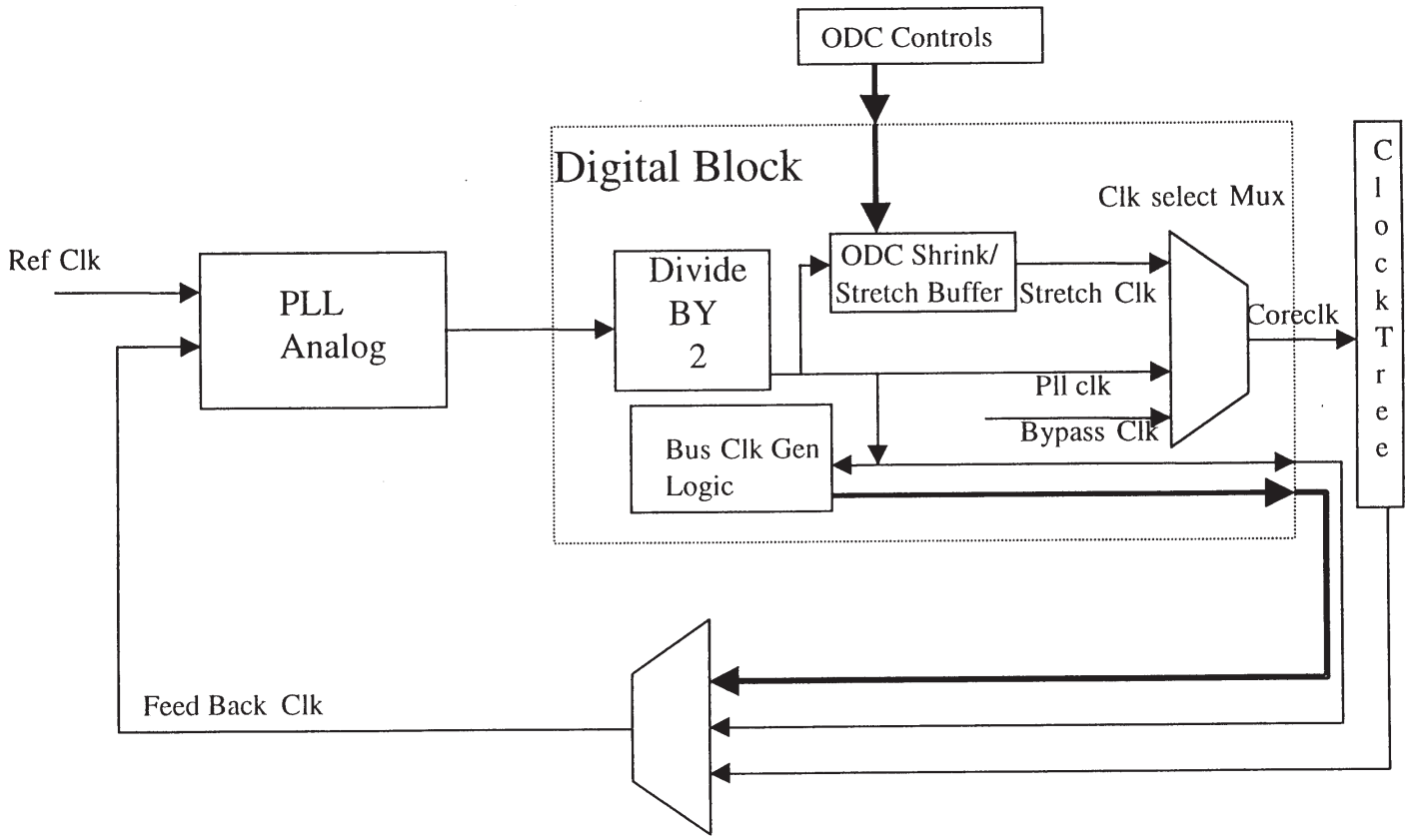

Figure 5.7.5: On-die-clock shrink/stretch functional block.

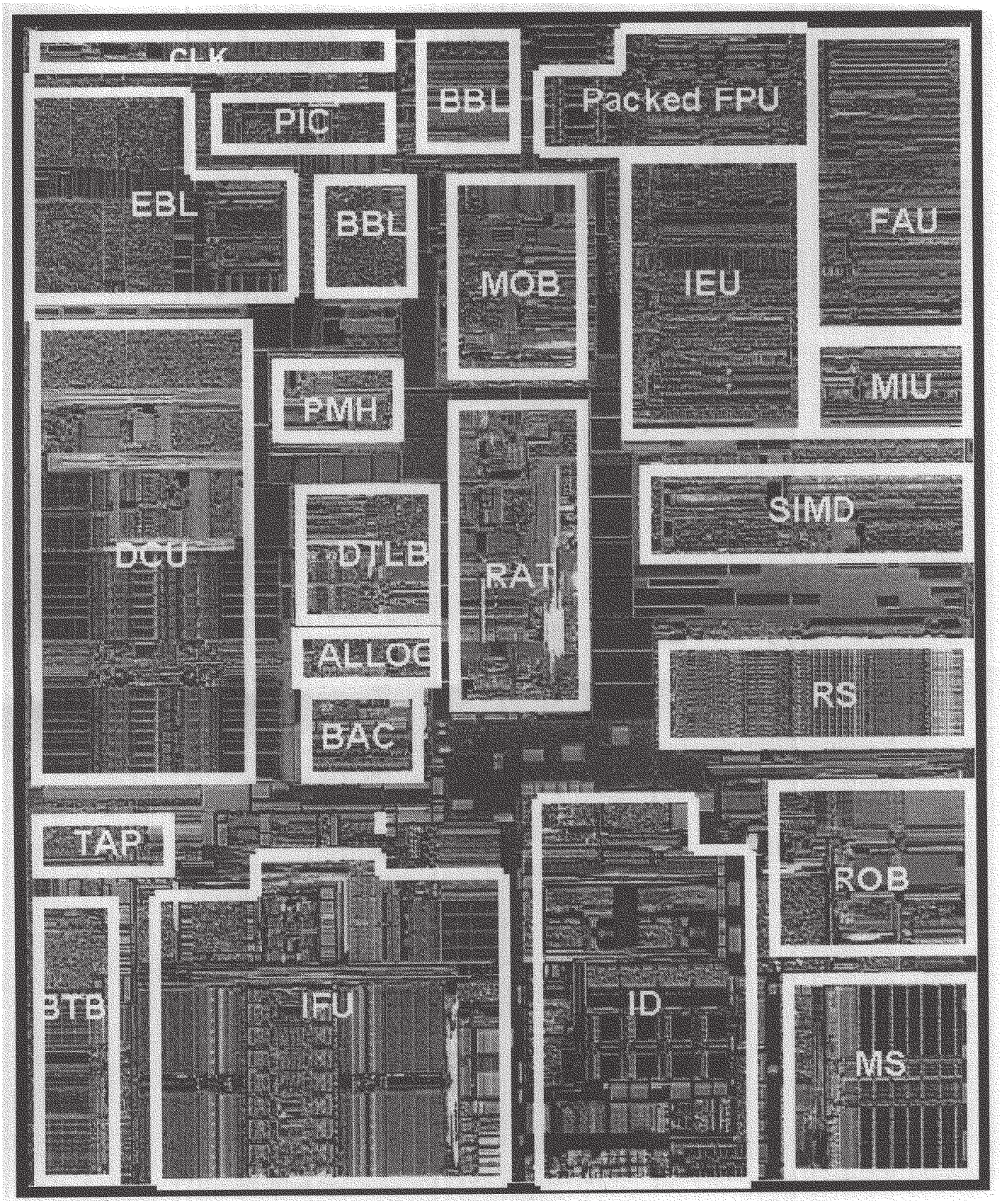

Figure 5.7.6: Die micrograph.

